Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 1, January 2015, pg.58 – 63

RESEARCH ARTICLE

STUDY OF FRAMEWORK OF PREDICTIVE DATA MINING FOR MEDICAL DATA

Mr. Rahul Pahlajani[1], Prof. Mr. Shrikant P. Akarte[2]

[1]ME (CSE) ,Second Year, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701

[2]Assistant Professor, Department of CSE, Prof. Ram Meghe Institute of Technology and Research, Badnera, Amravati. Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701

[1] rahulpahlajani@gmail.com, [2] s_akarte25@rediffmail.com

Abstract— In health organization Data Mining is one of the most causative fields of research that is become more and more popular. Data Mining plays an important role for bringing out new veers in healthcare organization which in turn useful for all the parties associated with this field. An important role is played by clinical prediction rules in medical practice. They hasten diagnosis and the unnecessary tests are limited. However, the process of rule creation is time consuming and costly. The automated rule induction from data supports the creation of clinical rules, with the current developments of efficient data mining algorithms and growing accessibility to medical data. This paper presents the depth and role of the research area of predictive data mining and to discuss available framework to cope with the problems of constructing, assessing and exploiting data mining models in clinical medicine.

Keywords— Data mining, medical data, Clinical prediction rules (CPRs), Diagnosis, Framework

---

I. INTRODUCTION

The term ‘data mining’ has been growingly used in the medical literature in the last few years. In general, the term has not been grounded to any exact definition but to some sort of common understanding of its meaning: the use of (new) methods and tools to analyse large amounts of data. In clinical medicine the aim of predictive data mining is to deduce models which can use patient specific information to predict the outcome of concern and to thereby support medical decision-making. For the construction of decision models for procedures such as medical prognosis, diagnosis and treatment planning Predictive data mining methods can be applied, which is verified and evaluated once and can be implanted within medical information systems.
The remainder of this paper is organized as follows. Section II provides a description of Related Work, which Section IV describes the Methodological Analysis of Predictive Data Mining which contains seven types of analysis. Section V describes Contribution Of Data Mining To Predictive Modelling In Clinical Medicines, section VI describes presents Comparative Analysis, finally section VII concludes this paper.

II. LITERATURE REVIEW

The process of selecting, exploring and modelling large amounts of data in order to expose unknown patterns or relationships which provide a clear and useful result to the data analyst is known as data mining [1]. The problems in Data mining are often solved by using pieces of different approaches drawn from computer science, including multi-dimensional

databases, machine learning, soft computing and data visualization, and from statistics, including hypothesis testing, clustering, classification and regression techniques. The trade of data mining lies in the appropriate choice and combination of these techniques to efficiently and reliably solve a given problem.

III. RELATED WORK

The term data mining has been mostly used by statisticians, data analysts, and the management information systems (MIS) communities. It has also achieved popularity in the field of database .In the mid-1990s, Fayyad et al. proposed the term data mining which is today become a synonym for 'Knowledge Discovery in Databases', emphasized the data analysis process rather than the use of specific analysis methods. In 1989 (Piatetsky-Shapiro 199t) introduces the KDD to emphasize that "knowledge" is the end product of a data-driven discovery. It has been popularized in artificial intelligence and machine learning.

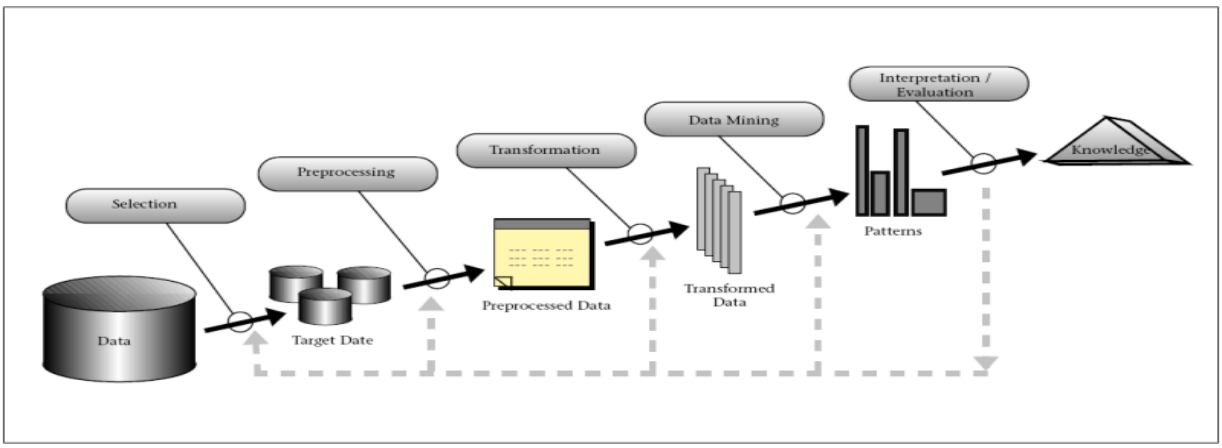

Figure 1. Stages of Knowledge Discovery Process.

IV. METHODOLOGICAL ANALYSIS OF PREDICTIVE DATA MINING

Predictive data mining methods derived from different research fields; they come in various feelings and may be compared on the basis of following factors,

- Their handling of missing data and noise;
- Their treatment of different types of attributes (categorical, ordinal, continuous);
- The reduction of the number of tests [2], that is, the reduction of attributes needed to derive the conclusion;
- Their ability to explain the decisions reached when models are used in decision-making, etc. the predictive data mining methods are studies and explained below.

A. Decision Trees.

The recursive data partitioning is used by decision trees, which induce transparent classifiers whose performance may suffer from data segmentation: the leaves in decision trees may include too few instances to obtain reliable predictions. Due to powerful heuristics the computational complexity of the induction algorithms is low[3, 4].

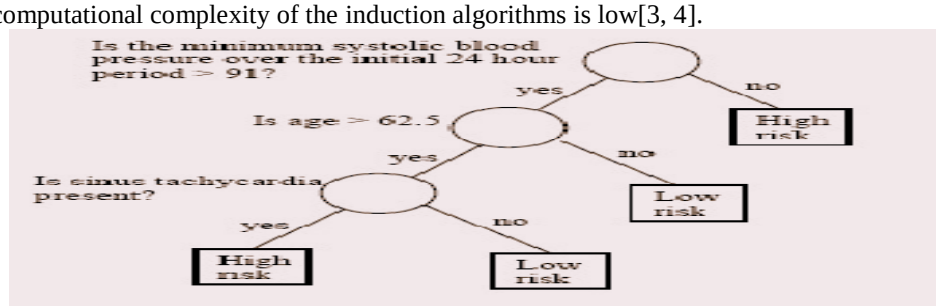

Figure 2. Classification by Decision Tree Induction

B. Decision Rules

It is in the form of 'IF condition-based-on attribute- values THEN outcome-value' may be constructed from induced decision trees as in the C4.5rules [3], or can be derived directly from the data as is the case with AQ and CN2 algorithms [5,6].Most of their characteristics with decision trees is shared by these algorithms in their performance. They can be computationally more expensive.

C. Logistic Regression

From statistics it is a powerful and well-established method [7]. It is an extension of ordinary regression and it can model a two-valued outcome which usually represents the occurrence or non- occurrence of some event. Like with the Na¨ıve Bayesian classifier, the underlying model for probability is multiplicative [8] but uses a more sophisticated method based on a maximum likelihood estimation to determine the coefficients in its probability formula. It is not straightforward to handel the missing attribute.
Nomogram represents the model very effectively[9,10].

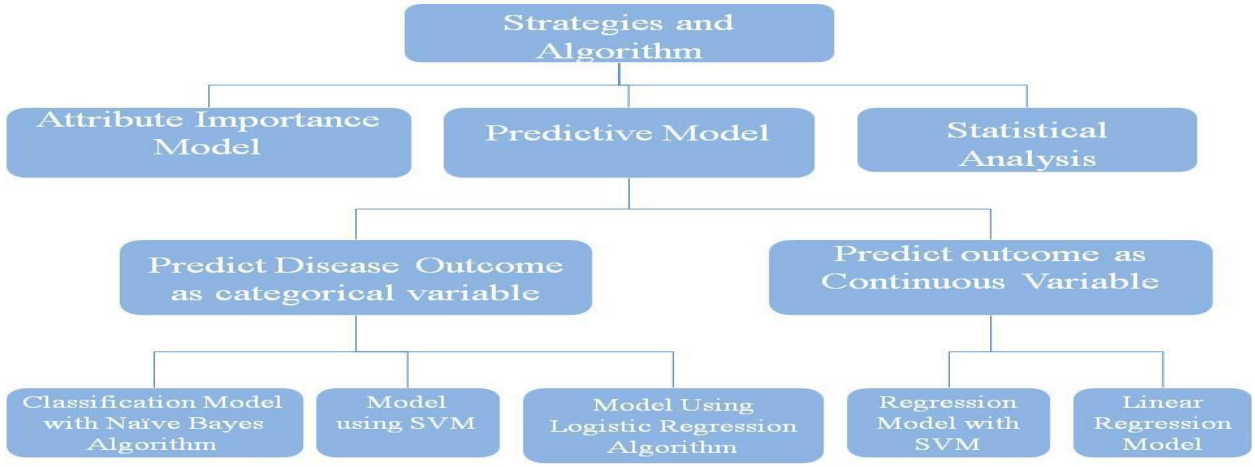

Figure 3. Functioning of Classification and Regression Techniques

D. Artificial Neural Networks

These networks were up until recently the most popular artificial intelligence-based data modeling algorithm used in clinical medicine. This is probably due to their good predictive performance, albeit they may have a number of deficiencies [11] which include high sensitivity to the parameters of the method—including those that determine the architecture of the network, initiation of the model whichever best – be hard to interpret by domain experts and high computational cost in training.

E. Support Vector Machines (SVM)

Today's most powerful classification algorithm is support vector machines in terms of predictive accuracy [12].They are based on strong mathematical foundations and statistical learning theory [13].

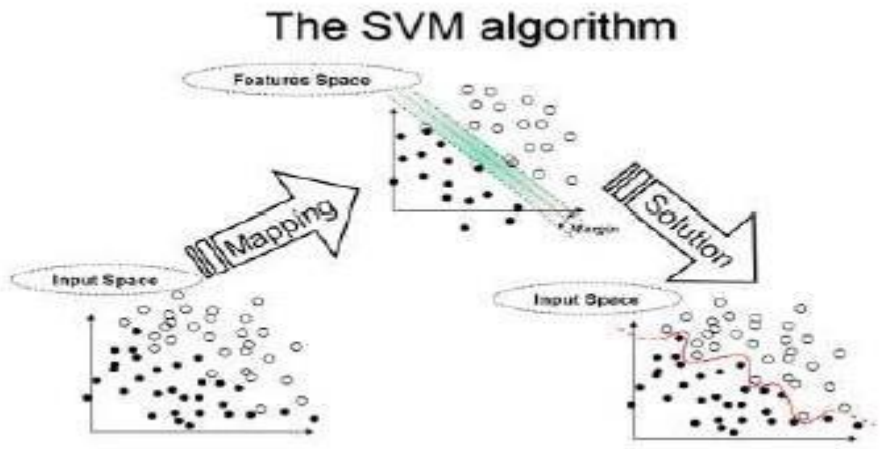

Figure 4. Support Vector Machine Classification

F. THE Na¨Ive Bayesian CLASSIFIER

The performance of na¨ıve Bayesian classifier is often at least comparable with other more sophisticated approaches [2, 14].

G. Bayesian Networks

These are the probabilistic graphical models that are able to conveniently express a joint probability distribution over a number of variables through a set of conditional probability distributions. A Bayesian network is a directed acyclic graph where each node represents a stochastic variable and a probabilistic dependency between a node and its parents is represented by arcs. Each variable *xi* is assumed to be independent of its non-descendants given its set of parents, pa(*xi*). Under this assumption, known as a Markov assumption, [15,16]the joint probability distribution of all variables (*x*) can be written following the so-called *chain rule*:

$$p(x) = \prod_{i=1}^{n} p(x_i|pa(x_i))$$

The algorithms for learning Bayesian networks from data are based on the framework of Bayesian model selection. The goal is to learn the structure *S* with the highest posterior probability distribution, given a data set *x*. Such a posterior probability distribution can be computed as:

$$p(S|x) = \frac{p(x|S)p(S)}{p(x)} \propto p(x|S)p(S)$$

H. The K-Nearest Neighbors Algorithm

The k-nearest neighbors algorithm is exalted by the approach often taken by domain experts who make decisions based on previously seen similar cases [8].The k-nearest neighbors classifier searches for the k most similar training instances and classifies based on their prevailing class For a given data instance.

V. CONTRIBUTION OF DATA MINING TO PREDICTIVE MODELING IN CLINICAL MEDICINE

In the clinical medicine the predictive models are proven themselves as tools for helping decision making that combine two or more items of patient data to predict clinical outcomes' [17].in several clinical circumstances Such models can be used by practitioners and may allow a prompt reaction to critical situations[18]. Data mining can contribute effectively in the development of clinically useful predictive models thanks to at least three inter-related aspects: (a) a comprehensive and purposive approach to data analysis that involves the application of methods and approaches drawn from different scientific areas; (b) the explanatory capability of such models; (c) the capability of using the background knowledge in the data analysis process.

VI. COMPARATIVE ANALYSIS

Medical data mining applications have several distinguishing features as Compared to data mining in business, marketing and the economy [19]. The most important one is that medicine is a safety critical context [20] in which explanations should support decision making activities. This means that the value of each result be higher than in other contexts: experiments can be costly due to the participation of the personnel and use of expensive instrumentation and due to the potential discomfort of the patients involved. In clinical mining, the data sets can be small and report non reproducible situations.

VII. CONCLUSION

The methods studied above are often an integral part of most modern data mining suites and, they are alone or in combination with pre-processing, often perform well and sufficiently fast. The biggest differences in the predictive performance and interpretability of results arise when treating clinical data. Throughout this review, concludes that both of these are important and if methods perform similarly with respect to accuracy those which offer an explanation and interpretable models should be preferred.

REFERENCES

[1] P. Giudici, Applied Data Mining Statistical Methods for Business and Industry, Wiley & Sons, 2003.

[2] N. Lavrac, I. Kononenko, E. Keravnou, M. Kukar, B. Zupan, Intelligent data analysis for medical diagnosis: using machine learning and temporal abstraction, AI Commun. 11 (1998) 191–218.

[3] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, Calif, 1993.

[4] L. Breiman, Classification and Regression Trees, Chapman & Hall, New York, London, 1993.

[5] P. Clark, T. Niblett, The CN2 Induction Algorithm, Mach. Learn. 3 (1989) 261–283.

[6] R.S. Michalski, K. Kaufman, Learning patterns in noisy data: the AQ approach, in: G. Paliouras, V.Karkaletsis,C.

[7] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd ed., Wiley, New York, 2000.

[8] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.

[9] J. Lubsen, J. Pool, E. van der Does, A practical device for the application of a diagnostic or prognostic function, Meth. Inf. Med. 17 (1978) 127–129.

[10] F.E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer, New York, 2001.

[11] G. Schwarzer, W. Vach, M. Schumacher, On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology, Stat. Med. 19 (2000) 541–561.

[12] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, New York, 2000.

[13] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[14] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artif. Intell. Med. 23 (2001) 89–109.

[15] E.H. Herskovits, J.P. Gerring, Application of a data-mining method based on Bayesian networks to lesion-deficit analysis, Neuroimage 19 (2003) 1664–1673.

[16] P. Sebastiani, M.F. Ramoni, V. Nolan, C.T. Baldwin, M.H. Steinberg, Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia, Nat. Genet. 37 (2005) 435–440.

[17] C.J. Wyatt, D.G. Altman, Prognostic models: clinically useful or quickly forgotten? BMJ (1995) 311.

[18] M.W. Kattan, M.J. Zelefsky, P.A. Kupelian, P.T. Scardino, Z. Fuks, S.A. Leibel, Pretreatment nomogram for predicting the outcome of three-dimensional conformal radiotherapy in prostate cancer, J. Clin. Oncol. 18 (2000) 3352–3359.

[19] K.J. Cios, G.W. Moore, Uniqueness of medical data mining, Artif. Intell. Med. 26 (2002) 1–24.

[20] J. Fox, S.K. Das, Safe and Sound: Artificial Intelligence In Hazardous Applications, MIT Press, Cambridge, Mass, 2000.

Regards:

Mr. Rahul Pahlajani[1], Prof. Mr. Shrikant P. Akarte[2]

[1]ME (CSE) ,Second Year,Department of CSE,Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701

[2]Assistant Professor, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati. Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

[1] rahulpahlajani@gmail.com, [2] s_akarte25@rediffmail.com, +91-8983622584,+91-9226792207.

